Lexical Semantic Techniques for Corpus Analysis

James Pustejovsky*
Brandeis University

Sabine Bergler†
Concordia University

Peter Anick‡
Digital Equipment Corporation

In this paper we outline a research program for computational linguistics, making extensive use of text corpora. We demonstrate how a semantic framework for lexical knowledge can suggest richer relationships among words in text beyond that of simple co-occurrence. The work suggests how linguistic phenomena such as metonymy and polysemy might be exploitable for semantic tagging of lexical items. Unlike with purely statistical collocational analyses, the framework of a semantic theory allows the automatic construction of predictions about deeper semantic relationships among words appearing in collocational systems. *We illustrate the approach for the acquisition of lexical information for several classes of nominals, and how such techniques can fine-tune the lexical structures acquired from an initial seeding of a machine-readable dictionary. In addition to conventional lexical semantic relations, we show how information concerning lexical presuppositions and preference relations can also be acquired from corpora, when analyzed with the appropriate semantic tools. Finally, we discuss the potential that corpus studies have for enriching the data set for theoretical linguistic research, as well as helping to confirm or disconfirm linguistic hypotheses.*

1. Introduction

The proliferation of on-line textual information poses an interesting challenge to linguistic researchers for several reasons. First, it provides the linguist with sentence and word usage information that has been difficult to collect and consequently largely ignored by linguists. Second, it has intensified the search for efficient automated indexing and retrieval techniques. Full-text indexing, in which all the content words in a document are used as keywords, is one of the most promising of recent automated approaches, yet its mediocre precision and recall characteristics indicate that there is much room for improvement (Croft 1989). The use of domain knowledge can enhance the effectiveness of a full-text system by providing related terms that can be used to broaden, narrow, or refocus a query at retrieval time (Debili, Fluhr, and Radasua 1988; Anick et al. 1989. Likewise, domain knowledge may be applied at indexing time to do word sense disambiguation (Krovetz and Croft 1989) or content analysis (Jacobs 1991). Unfortunately, for many domains, such knowledge, even in the form of a thesaurus, is either not available or is incomplete with respect to the vocabulary of the texts indexed.

* Computer Science Department, Brandeis University, Waltham MA 02254.
† Computer Science Department, Concordia University, Montreal, Quebec H3G 1M8, Canada.
‡ Digital Equipment Corporation, 111 Locke Drive LM02-1/D12, Marlboro MA 01752.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 1993		**00-00-1993 to 00-00-1993**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Lexical Semantic Techniques for Corpus Analysis	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Brandeis University ,Department of Computer Science,Waltham,MA,02254**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**28**	

In this paper we examine how linguistic phenomena such as metonymy and polysemy might be exploited for the semantic tagging of lexical items. Unlike purely statistical collocational analyses, employing a semantic theory allows for the automatic construction of deeper semantic relationships among words appearing in collocational systems. We illustrate the approach for the acquisition of lexical information for several classes of nominals, and how such techniques can fine-tune the lexical structures acquired from an initial seeding of a machine-readable dictionary. In addition to conventional lexical semantic relations, we show how information concerning lexical presuppositions and preference relations (Wilks 1978) can also be acquired from corpora, when analyzed with the appropriate semantic tools. Finally, we discuss the potential that corpus studies have for enriching the data set for theoretical linguistic research, as well as helping to confirm or disconfirm linguistic hypotheses.

The aim of our research is to discover what kinds of knowledge can be reliably acquired through the use of these methods, exploiting, as they do, general linguistic knowledge rather than domain knowledge. In this respect, our program is similar to Zernik's (1989) work on extracting verb semantics from corpora using lexical categories. Our research, however, differs in two respects: first, we employ a more expressive lexical semantics; second, our focus is on all major categories in the language, and not just verbs. This is important since for full-text information retrieval, information about nominals is paramount, as most queries tend to be expressed as conjunctions of nouns. From a theoretical perspective, we believe that the contribution of the lexical semantics of nominals to the overall structure of the lexicon has been somewhat neglected, relative to that of verbs. While Zernik (1989) presents ambiguity and metonymy as a potential obstacle to effective corpus analysis, we believe that the existence of motivated metonymic structures actually provides valuable clues for semantic analysis of nouns in a corpus.

We will assume, for this paper, the general framework of a generative lexicon as outlined in Pustejovsky (1991). In particular, we make use of the principles of type coercion and qualia structure. This model of semantic knowledge associated with words is based on a system of generative devices that is able to recursively define new word senses for lexical items in the language. These devices and the associated dictionary make up a *generative lexicon*, where semantic information is distributed throughout the lexicon to all categories. The general framework assumes four basic levels of semantic description: argument structure, qualia structure, lexical inheritance structure, and event structure.

Connecting these different levels is a set of generative devices that provide for the compositional interpretation of words in context. The most important of these devices is a semantic transformation called *type coercion*—analogous to coercion in programming languages—which captures the semantic relatedness between syntactically distinct expressions. As an operation on types within a λ-calculus, type coercion can be seen as transforming a monomorphic language into one with polymorphic types (cf. Cardelli and Wegner 1985). Argument, event, and qualia types must conform to the well-formedness conditions defined by the type system defined by the lexical inheritance structure when undergoing operations of semantic composition.[1]

1 The details of type coercion need not concern us here. Briefly, however, whenever there exists a grammatical environment where more than one syntactic type satisfies the semantic type selected by the governing element, the governing element can be analyzed as coercing a range of surface types into a single semantic type. An example of *subject type coercion* is a causative verb, semantically selecting an event as subject (as in (i)), but syntactically permitting a nonevent denoting NP (as in (ii)):

i. The flood killed the grass.

ii. The herbicide killed the grass.

One component of this approach, the qualia structure, specifies the different aspects of a word's meaning through the use of subtyping. These include the subtypes CONSTITUTIVE, FORMAL, TELIC, and AGENTIVE. To illustrate how these are used, the qualia structure for *book* is given below.[2]

$$\begin{bmatrix} \textbf{book(x,y)} \\ \text{CONST} = \textbf{information(y)} \\ \text{FORMAL} = \textbf{physobj(x)} \\ \text{TELIC} = \textbf{read(T,w,y)} \\ \text{AGENTIVE} = \textbf{write(T,z,y)} \end{bmatrix}$$

This structured representation allows one to use the same lexical entry in different contexts, where the word refers to different qualia of the noun's denotation. For example, the sentences in (1)–(3) below refer to different aspects (or *qualia*) of the general meaning of *book*.[3]

Example 1
This book weighs four ounces.

Example 2
John finished a book.

Example 3
This is an interesting book.

Example 1 makes reference to the formal role, while 3 refers to the constitutive role. Example 2, however, can refer to either the telic or the agentive aspects given above. The utility of such knowledge for information retrieval is readily apparent. This theory claims that noun meanings should make reference to related concepts and the relations into which they enter. The qualia structure, thus, can be viewed as a kind of generic template for structuring this knowledge. Such information about how nouns relate to other lexical items and their concepts might prove to be much more useful in full-text information retrieval than what has come from standard statistical techniques.

To illustrate how such semantic structuring might be useful, consider the general class of artifact nouns. A generative view of the lexicon predicts that by classifying an element into a particular category, we can generate many aspects of its semantic structure, and hence, its syntactic behavior. For example, the representation above for *book* refers to several word senses, all of which are logically related by the semantic template for an artifactual object. That is, it contains information, it has a material extension, it serves some function, and it is created by some particular act or event.

---

2 Briefly, the qualia can be defined as follows:

- CONSTITUTIVE: the relation between an object and its constituent parts;
- FORMAL: that which distinguishes it within a larger domain;
- TELIC: its purpose and function;
- AGENTIVE: factors involved in its origin or "bringing it about."

In the qualia structures given below, we adopt the convention that $[\alpha, \beta]$ denotes conjunction of formulas within the feature structure, while $[\alpha; \beta]$ will denote disjunction.

3 A related approach for expressing the different semantic relations of nominals in distinguished contexts is given in Bierwisch (1983).

Such an analysis allows us to minimally structure objects according to these four qualia.

As an example of how objects cluster according to these dimensions, we will briefly consider three object types: (1) containers (of information), e.g., *book, tape, record*; (2) instruments, e.g., *gun, hammer, paintbrush*; and (3) figure-ground objects, e.g., *door, room, fireplace*. Because of how their qualia structures differ, these classes appear in vastly different grammatical contexts.

As with containers in general, information containers permit metonymic extensions between the container and the material contained within it. Collocations such as those in Examples 4 through 7 indicate that this metonymy is grammaticalized through specific and systematic head-PP constructions.

Example 4
read a book

Example 5
read a story in a book

Example 6
read a tape

Example 7
read the information on the tape

Instruments, on the other hand, display classic agent–instrument causative alternations, such as those in Examples 8 through 11 (cf. Fillmore 1968; Lakoff 1968, 1970).

Example 8
...smash the vase with the hammer

Example 9
The hammer smashed the vase.

Example 10
...kill him with a gun

Example 11
The gun killed him.

Finally, *figure-ground nominals* (Pustejovsky and Anick 1988) permit perspective shifts such as those in Examples 12 through 15. These are nouns that refer to physical objects as well as the specific enclosure or aperture associated with it.

Example 12
John painted the door.

Example 13
John walked through the door.

Example 14
John is scrubbing the fireplace.

Example 15
The smoke filled the fireplace.

That is, *paint* and *scrub* are actions on physical objects while *walk through* and *fill* are processes in spaces. These collocational patterns, we argue, are systematically predictable from the lexical semantics of the noun, and we term such sets of collocated structures *lexical conceptual paradigms (LCPs).*[4]

To make this point clearer, let us consider a specific example of an LCP from the computer science domain, namely for the noun *tape*. Because of the particular metonymy observed for a noun like *tape*, we will classify it as belonging to the container/containee LCP. This general class is represented as follows, where P and Q are predicate variables:[5]

$$\begin{bmatrix} \textbf{container(x,y)} \\ \text{CONST} = \textbf{P(y)} \\ \text{FORMAL} = \textbf{Q(x)} \\ \text{TELIC} = \textbf{hold(S,x,y)} \end{bmatrix}$$

The LCP is a generic qualia structure that captures not only the semantic relationship between arguments types of a relation, but also, through corpus-tuning, the collocation relations that realize these roles. The telic function of a container, for example, is the relation `hold`, but this underspecifies which spatial prepositions would adequately satisfy this semantic relation (e.g. *in, on, inside*, etc.).

In this view, a noun such as *tape* would have the following qualia structure:

$$\begin{bmatrix} \textbf{tape(x,y)} \\ \text{CONST} = \textbf{information(y)} \\ \text{FORMAL} = \textbf{physobj(x),2-dimen(x)} \\ \text{TELIC} = \textbf{contain(S,x,y)} \\ \text{AGENTIVE} = \textbf{write(T,z,y)} \end{bmatrix}$$

This states that a tape is an "information container" that is also a two-dimensional physical object, where the information is written onto the object.[6] With such nouns, a *logical metonymy* exists (as the result of type coercion), when the logical argument of a semantic type, which is selected by a function of some sort, denotes the semantic type itself. Thus, in this example, the type selected for by a verb such as *read* refers to the "information" argument for *tape*, while a verb such as *carry* would select for the "physical object" argument. They are, however, logically related, since the noun itself denotes a relation.

The representation above simply states that any semantics for *tape* must logically make reference to the object itself (formal), what it can contain (const), what purpose

4 This relates to Mel'čuk's lexical functions and the syntactic structures they associate with an element. See Mel'čuk (1988) and references therein. Cruse (1986, 1992) and Nunberg (1978) discuss the foregrounding and backgrounding of information with respect to similar examples.

5 Within the qualia structure for a term, FORMAL and CONST roles typically refer to the object domain while TELIC and AGENTIVE refer to events. Hence, the first parameter in the latter two roles refers to an event sort, i.e., a state (S), process (P), or transition (T).

6 The appropriate selection of a surface spatial preposition will follow from its formal type specification as a `2-dimen` object. Cf. Pustejovsky (in press) for details.

it serves (telic), and how it arises (agentive). This provides us with a semantic representation that can capture the multiple perspectives a single lexical item may assume in different contexts. Yet, the qualia for a lexical item such as *tape* are not isolated values for that one word, but are integrated into a global knowledge base indicating how these senses relate to other lexical items and their senses. This is the contribution of inheritance and the hierarchical structuring of knowledge (cf. Evans and Gazdar 1990; Copestake and Briscoe 1992; Russell et al. 1992). In Pustejovsky (1991) it is suggested that there are two types of relational structures for lexical knowledge; a *fixed* inheritance similar to that of an `is-a` hierarchy (cf. Touretzky 1986); and a dynamic structure that operates *generatively* from the qualia structure of a lexical item to create a relational structure for ad hoc categories.[7]

Reviewing briefly, the basic idea is that semantics allows for the dynamic creation of arbitrary concepts through the application of certain transformations to lexical meanings. Thus for every predicate, Q, we can generate its opposition, $\neg Q$. Similarly, these two predicates can be related temporally to generate the transition events defining this opposition. These operations include but may not be limited to: $\neg$, negation; $\leq$, temporal precedence; $\geq$, temporal succession; $=$, temporal equivalence; and *act*, an operator adding agency to an argument. We will call the concept space generated by these operations the *Projective Conclusion Space* of a specific quale for a lexical item. To return to the example of *tape* above, the predicates *read* and *copy* are related to the telic value by just such an operation, while predicates such as *mount* and *dismount*—i.e. *unmount*—are related to the formal role. Following the previous discussion, with *mounted* as the predicate Q, successive applications of the negation and temporal precedence operators derives the transition verbs *mount* and *dismount*.[8] We return to a discussion of this in Section 3, and to how this space relates to statistically significant collocations in text.

It is our view that the approach outlined above for representing lexical knowledge can be put to use in the service of information retrieval tasks. In this respect, our proposal can be compared to attempts at object classification in information science. One approach, known as *faceted classification* (Vickery 1975) proceeds roughly as follows: collect all terms lying within a field; then group the terms into facets by assigning them to categories. Typical examples of this are *state*, *property*, *reaction*, and *device*. However, each subject area is likely to have its own sets of categories, which makes it difficult to re-use a set of facet classifications.[9]

Even if the relational information provided by the qualia structure and inheritance would improve performance in information retrieval tasks, one problem still remains, namely that it would be very time-consuming to hand-code such structures for all nouns in a domain. Since it is our belief that such representations are generic structures across all domains, it is our long-term goal to develop methods for automatically extracting these relations and values from on-line corpora. In the sections that follow, we describe several experiments indicating that the qualia structures do, in fact, correlate with well-behaved collocational patterns, thereby allowing us to perform structure-matching operations over corpora to find these relations.

7 This is similar to thesauruslike structures, within the IR community, cf. for example Sparck Jones (1981).

8 Details of the derivation are as follows. Let Q be *mounted*, then $\neg Q$ gives $\neg$*mounted*, and $\leq$ applied to these two states gives $Q \leq \neg Q$, which is lexicalized as *dismount*. A similar derivation exists for *mount*. Cf. Pustejovsky (1991) for details.

9 This is reflected in the sublanguage work of Grishman, Hirschman, and Nhan (1986), whose automated discovery procedures are aimed at clustering nouns into categories like *diagnosis* and *symptom*.

2. Seeding Lexical Structures from MRDs

In this section we discuss briefly how a lexical semantic theory can help in extracting information from machine-readable dictionaries (MRDs). We describe research on conversion of a machine-tractable dictionary (Wilks et al. 1993) into a usable lexical knowledge base (Boguraev 1991). Although the results here are preliminary, it is important to mention the process of converting an MRD into a lexical knowledge base, so that the process of corpus-tuning is put into the proper perspective. The initial seeding of lexical structures is being done independently both from the Oxford Advanced Learners Dictionary (OALD) and from lexical entries in the Longman Dictionary of Contemporary English (Procter, Ilson, and Ayto 1978). These are then automatically adapted to the format of generative lexical structures. It is these lexical structures that are then statistically tuned against the corpus, following the methods outlined in Anick and Pustejovsky (1990) and Pustejovsky (1992).

Previous work by Amsler (1980), Calzolari (1984), Chodorow, Byrd, and Heidorn (1985), Byrd et al. (1987), Markowitz, Ahlswede, and Evens (1986), and Nakamura and Nagao (1988) showed that taxonomic information and certain semantic relations can be extracted from MRDs using fairly simple techniques. Later work by Veronis and Ide (1991), Klavans, Chodorow, and Wacholder (1990), and Wilks et al. (1992) provides us with a number of techniques for transfering information from MRDs to a representation language such as that described in the previous section. Our goal is to automate, to the extent possible, the initial construction of these structures.

Extensive research has been done on the kind of information needed by natural language programs and on the representation of that information (Wang, Vandendorpe, and Evens 1985; Ahlswede and Evens 1988). Following Boguraev et al. (1989) and Wilks et al. of 1989), we believe that much of what is needed for NLP lexicons can be found either explicitly or implicitly in a dictionary, and empirical evidence suggests that this information gives rise to a sufficiently rich lexical representation for use in extracting information from texts. Techniques for identifying explicit information in machine-readable dictionaries have been developed by many researchers (Boguraev et al. 1989; Slator 1988; Slator and Wilks 1987; Guthrie et al. 1990) and are well understood. Many properties of a word sense or the semantic relationships between word senses are available in MRDs, but this information can only be identified computationally through some analysis of the definition text of an entry (Atkins 1991). Some research has already been done in this area. Alshawi (1987), Boguraev et al. (1989), Vossen, Meijs, and den Broeder (1989), and the work described in Wilks et al. (1992) have made explicit some kinds of implicit information found in MRDs. Here we propose to refine and merge some of the previous techniques to make explicit the implicit information specified by a theory of generative lexicons.

Given what we described above for the lexical structures for nominals, we can identify these semantic relations in the OALD and LDOCE by pattern matching on the parse trees of definitions. To illustrate what specific information can be derived by automatic seeding from machine-readable dictionaries, consider the following examples.[10] For example, the LDOCE definition for *book* is:

> "a collection of sheets of paper fastened together as a thing to be read, or to be written in"

10 The following lexical entries, termed *gls's*, are taken from the lexical databases derived from the OALD using tools developed by Peter Dilworth, and from LDOCE using a combination of tools developed by Louise Guthrie, Gees Stein, and Pete Dilworth.

while the OALD provides a somewhat different definition:

> "number of sheet of papers, either printed or blank, fastened together in a cover."

Note that both definitions are close to, but not identical to the information structure suggested in the previous section, using a qualia structure for nominals. LDOCE suggests *write in* rather than *write* as the value for the telic role, while the OALD suggests nothing for this role. Furthermore, although the physical contents of a book as "a collection of sheets of paper" is mentioned, nowhere is *information* made reference to in the definition. When the dictionary fails to provide the value for a semantic role, the information must be either hand-entered or the lexical structure must be tuned against a large corpus, in the hope of extracting such features automatically. We turn to this issue in the next two sections.

Although the two dictionaries differ in substantial respects, it is remarkable how systematic the definition structures are for extracting semantic information, if there is a clear idea how this information should be structured. For example, from the following OALD definition for *cigarette,*

> `cigarette` *n* roll of shredded tobacco enclosed in thin paper for smoking.

the initial lexical structure below is generated.

```
gls(cigarette,
   syn([type(n),
        code(C)]),
   qualia([formal([roll]),
           telic([smoking]),
           const([tobacco,paper]),
           agent([enclosed])]),
   cospec([])).
```

Parsing the LDOCE entry for the same noun results in a different lexical structure:

> `cigarette` *n* finely cut shredded tobacco rolled in a narrow tube of thin paper for smoking.

```
gls(cigarette,
   syn([type(n),
        code(C),
        ldoce_id(cigarette_0_1)]),
   qualia([formal([tube]),
           telic([smoking]),
           const([tobacco,paper]),
           agent([rolled])]),
   cospec([])).
```

One obvious problem with the above representation is that there is no information indicating how the word being defined binds to the relations in the qualia. Currently,

subsequent routines providing for argument binding analyze the relational structure for particular aspects of noun meaning, giving us a lexical structure fairly close to what we need for representation and retrieval purposes, although the result is in no way ideal or uniform over all nominal forms. (cf. Cowie, Guthrie, and Pustejovsky [1992] for details of this operation on LDOCE.):[11]

$$\begin{bmatrix} \textbf{cigarette(x)} \\ \text{CONST} = \textbf{tobacco(y),shredded(y),paper(z)} \\ \text{FORMAL} = \textbf{roll(x)} \\ \text{TELIC} = \textbf{smoke(T,w,x)} \\ \text{AGENTIVE} = \textbf{artifact(x)} \end{bmatrix}$$

In a related set of experiments performed while constructing a large lexical database for data extraction purposes, we seeded a lexicon with 6000 verbs from LDOCE. This process and the corpus tuning for both argument typing and subcategorization acquisition are described in Cowie, Guthrie, and Pustejovsky (1992) and Pustejovsky et al. (1992).

In summary, based on a theory of lexical semantics, we have discussed how an MRD can be useful as a corpus for automatically seeding lexical structures. Rather than addressing the specific problems inherent in converting MRDs into useful lexicons, we have emphasized how it provides us, in a sense, with a generic vocabulary from which to begin lexical acquisition over corpora. In the next section, we will address the problem of taking these initial, and often very incomplete lexical structures, and enriching them with information acquired from corpus analysis. As mentioned in the previous section, the power of a generative lexicon is that it takes much of the burden of semantic interpretation off of the verbal system by supplying a much richer semantics for nouns and adjectives. This makes the lexical structures ideal as an initial representation for knowledge acquisition and subsequent information retrieval tasks.

3. Knowledge Acquisition from Corpora

A machine-readable dictionary provides the raw material from which to construct computationally useful representations of the generic vocabulary contained within it. The lexical structures discussed in the previous section are one example of how such information can be exploited. Many sublanguages, however, are poorly represented in on-line dictionaries, if represented at all. Vocabularies geared to specialized domains will be necessary for many applications, such as text categorization and information retrieval. The second area of our research program that we discuss is aimed at developing techniques for building sublanguage lexicons via syntactic and statistical corpus analysis coupled with analytic techniques based on the tenets of generative lexicon theory.

To understand fully the experiments described in the next two sections, we will refer to several semantic notions introduced in previous sections. These include *type coercion*, where a lexical item requires a specific type specification for its argument, and

11 As one reviewer correctly pointed out, more than simple argument binding is involved here. For example, the model must know that paper can enclose shredded tobacco, but not the reverse. Such information, typically part of commonsense knowledge, is well outside the domain of lexical semantics, as envisioned here. One approach to this problem, consistent with our methodology, is to examine the corpus and the collocations that result from training on specific qualia relations. Further work will hopefully clarify the nature of this problem, and whether it is best treated lexically or not.

the argument is able to change type accordingly—this explains the behavior of logical metonymy and the syntactic variation seen in complements to verbs and nominals; and *cospecification*, a semantic tagging of what collocational patterns the lexical item may enter into.

Metonymy, in this view, can be seen as a case of the "licensed violation" of selectional restrictions. For example, while the verb *announce* selects for a human subject, sentences like *The Dow Corporation announced third quarter losses* are not only an acceptable paraphrase of the selectionally correct form *Mr. Dow Jr. announced third quarter losses for Dow Corp*, but they are the preferred form in the corpora being examined. This is an example of subject *type coercion*, where the semantics for Dow Corp as a company must specify that there is a human typically associated with such official pronouncements (see Section 5).[12]

For one set of experiments, we used a corpus of approximately 3,000 articles written by Digital Equipment Corporation's Customer Support Specialists for an on-line computer troubleshooting library. The articles, each one- to two-page long descriptions of a problem and its solution, comprise about 1 million words. Our analysis proceeds in two phases. In the first phase, we pre-process the corpus to build a database of phrasal relationships. This consists briefly of the following steps:

1. **Perform unknown word resolution to the corpus.** The corpus is searched for strings that are not members of a 25,000 word generic on-line English lexicon. Morphological analysis is then applied to these unknown strings to identify candidate citation forms and their likely morphological paradigms. Unless morphological evidence indicates otherwise, we enter unknown words into the lexicon as regular nouns; if there is evidence of some other morphological paradigm, such as verbal or adjectival suffixes, the word is entered into the lexicon accordingly.

2. **Corpus tagging.** Once the lexicon is updated to include the new single word forms in the domain, the corpus is tagged with part-of-speech indicators. Any words that are ambiguous with respect to category are disambiguated according to a set of several dozen ordered disambiguation heuristics, which choose a category based on the categories of the words immediately preceding and following the ambiguous term.

3. **Partial parsing.** The tagged corpus is then segmented into a flat sequence of phrasal groupings, using closed class words such as prepositions and determiners, as well as certain part-of-speech transitions, to indicate likely phrase boundaries. No attempt is made to construct a full parse tree or resolve prepositional phrase attachment, conjunction scoping, etc. A concordance is constructed, identifying, for each word appearing in the corpus, the set of sentences, phrases, and phrase locations in which the word appears.

12 Within the current framework, a distinction is made between *logical metonymy*, where the metonymic extension or relation is transparent from the lexical semantics of the coerced phrase, and *conventional metonymy*, where the relation may not be directly calculated from information provided grammatically. For example, in the sentence "The Boston office called today," it is not clear from logical metonymy what relation *Boston* bears to *office* other than location; i.e., it is not obvious that it is a branch office. This is well beyond lexical semantics (cf. Lakoff 1987 and Martin 1990).

The database of partially parsed sentences provides the raw material for a number of sublanguage analyses. This begins the second phase of analysis:

1. **Noun compound recognition and bracketing.** In technical sublanguages, noun compounds are often employed to expand the working vocabulary without the invention of new word forms. It is therefore useful in applications such as lexicon-assisted full-text information retrieval (Anick 1992) to include such noun compounds as lexical items for both querying and thesaurus browsing. We construct bracketed noun compounds from our database of partial parses in a two-step process. The first simply searches the corpus for (recurring) contiguous sequences of nouns. Then, to bracket each compound that includes more than two nouns, we test whether possible subcomponents of the phrase exist on their own (as complete noun compounds) elsewhere in the corpus. Sample bracketed compounds derived from the computer troubleshooting database include `[[system management] utility]`, `[TK50 [tape drive]]`, `[[database management] system]`.

2. **Generation of taxonomic relationships** on the basis of collocational information. Technical sublanguages often express subclass relationships in noun compounds of the form `<instance-name> <class-name>`, as in "Unix operating system" and "C language." Unfortunately, noun compounds are also employed to express numerous other relationships, as in "Unix kernel" and "C debugger." We have found, however, that collocational evidence can be employed to suggest which noun compounds reflect taxonomic relationships, using a strategy similar to that employed by Hindle (1990) for detecting synonyms. Given a term T, we extract from the phrase database those nouns N_i that appear as the head of any phrase in which T is the immediately preceding term. These nouns represent candidate classes of which T may be a member. We then generate the set of verbs that take T as direct object and calculate the mutual information value for each *verb*/T collocation (cf. Hindle 1990). We do the same for each noun N_i. Under the assumption that instance and class nouns are likely to co-occur with the same verbs, we compute a similarity score between T and each noun N_i, by summing the product of the mutual information values for those verbs occurring with both nouns. (Verbs with negative mutual information values are left out of the calculation.) The noun with the highest similarity score is often the class of which T is an instance, as illustrated by the sample results in Figure 1. For each word displayed in Figure 1, its "class" is the head noun with the highest similarity score. Other head nouns occurring with the word as modifier are listed as well.

   As with all the automated procedures described here, this algorithm yields useful, but imperfect results. The class chosen for "VMS," for example, is incorrect, and may reflect the fact that in a DEC troubleshooting database, authors see no need to further specify VMS as "VMS operating system." A more interesting observation is that, among the collocations associated with the terms, there are often several that might qualify as classes of which the term is an instance, e.g., DECWindows could also be classified as "software"; TK50 might also qualify as "tape." From a generative lexicon perspective, these alternative classifications reflect multiple inheritance through the noun's

word	class	score	other collocations
HSC	controller	27.69	device, disk, path, message
BACKUP	operation	34.18	disk, tape, process, saveset
RL02	disk	15.93	media, kit, pack
TK50	cartridge	39.17	tape, kit, density, format
ACCVIO	error	14.35	problem
VAX	product	23.28	configuration, node, editor, hardware
VMS	support	7.98	product, upgrade, installation
upgrade	procedure	12.27	phase, option, support, prerequisite
DCL	level	9.14	command, procedure, access, error
CHECKSUM	value	4.45	character, operation, error
EDT	editor	11.58	session, conversion, search, problem
TPU	command	3.62	editor, session, function, debugger
RTL	error	1.58	routine, library
DECWindows	environment	75.46	image, application, intrinsics, software

Figure 1
Classification of nouns from a computer troubleshooting corpus.

qualia. That is, "cartridge" is further specifying the formal role of tape for TK50. DECWindows is functionally an "environment," its telic role, while "software" characterizes its formal quale.

3. **Extraction of information relating to noun's qualia.** Under certain circumstances, it may be possible to elicit information about a noun's qualia from automated procedures on a corpus. In this line of research, we haved employed the notion of "lexical conceptual paradigm" described above. An LCP relates a set of syntactic behaviors to the lexical semantic structures of the participating lexical items.

   For example, the set of expressions involving the word "tape" in the context of its use as a secondary storage device suggests that it fits the *container artifact* schema of the qualia structure, with "information" and "file" as its containees:

   (a) *read information from tape*
   (b) *write file to tape*
   (c) *read information on tape*
   (d) *read tape*
   (e) *write tape*

   As mentioned in Section 1, containers tend to appear as objects of the prepositions *to, from, in,* and *on* as well as in direct object position, in which case they are typically serving metonymically for the containee. Thus, the container LCP relates the set of generalized syntactic patterns

   $V_i\ N_j\ \{\texttt{to}, \texttt{from}, \texttt{on}\}\ N_k$
   $V_i\ N_j$
   $V_i\ N_k$

   to the underlying lexical semantic structure given below.

$$\begin{bmatrix} \textbf{container(x,y)} \\ \text{CONST} = \textbf{P(y)} \\ \text{FORMAL} = \textbf{Q(x)} \\ \text{TELIC} = \textbf{hold(S,x,y)} \end{bmatrix}$$

verb	MI	count
unload	5.43	5
position	3.92	5
mount	3.77	29
initialize	3.18	10
dismount	2.88	5
read	1.40	7
load	1.18	4
restore	0.80	3
write	−0.24	2
copy	−2.55	1

Figure 2
Verbs associated with direct object tape as direct object.

This LCP includes a nominal alternation between the container and containee in the object position of verbs. For *tape*, this alternation is manifested for verbs that predicate the telic role of data storage but not the formal role of physical object, which refers to the object as a whole regardless of its contents:

- TELIC = data-storage
  (a) *read* (tape/data from tape)
  (b) *write* (tape/data on tape)
  (c) *copy* (tape/data from tape)
- FORMAL= physical object
  (a) *mount* (tape)
  (b) *dismount* (tape)

We have explored the use of heuristics to distinguish those predicates that relate to the *Telic* quale of the noun. Consider the word *tape*, which occurs as the direct object in 107 sentences in our corpus. It appears with a total of 34 different verbs. By applying the mutual information metric (MI) to the verb–object pairs, we can sort the verbs accordingly, giving us the table of verbs most highly associated with *tape*, shown in Figure 2. While the mutual information statistic does a good job of identifying verbs that semantically relate to the word *tape*, it provides no information about how the verbs relate to the noun's qualia structure. That is, verbs such as *unload*, *position*, and *mount* are selecting for the formal quale of *tape*, a physical object that can be physically manipulated with respect to a tape drive. *Read*, *write*, and *copy*, on the other hand, relate to the telic role, the function of a tape as a medium for storing information.

Our hypothesis was that the nominal alternation can help to distinguish the two sets of verbs. We reasoned that, if the alternation is based on the container/containee metonymy, then it will be those verbs that apply to the telic role of the direct object that participate in the alternation. We tested this hypothesis as follows.

We generated a candidate set of containees for *tape* by identifying all the nouns that appeared in the corpus to the left of the adjunct *on tape*.

$\mathcal{S}_1$	Verbs with `tape` as object
$\mathcal{S}_2$	Verbs with a containee of `tape` as object
$\mathcal{S}_1 \cap \mathcal{S}_2$	{restore, create, write, read, copy, replace}
$\mathcal{S}_1 - \mathcal{S}_2$	{mount, initialize, unload, position, dismount, load, allocate }
$\mathcal{S}_1$	Verbs with `disk` as object
$\mathcal{S}_2$	Verbs with a containee of `disk` as object
$\mathcal{S}_1 \cap \mathcal{S}_2$	{compress, restore, disable, rebuild, modify, recover, search, copy}
$\mathcal{S}_1 - \mathcal{S}_2$	{mount, initialize, boot, dismount, serve, }
$\mathcal{S}_1$	Verbs with `directory` as object
$\mathcal{S}_2$	Verbs with containee of `directory` as object
$\mathcal{S}_1 \cap \mathcal{S}_2$	{create, recreate, delete, store, rename, check}
$\mathcal{S}_1 - \mathcal{S}_2$	{own, miss, search, review}

Figure 3
Intersection and set difference for three container nouns.

Then we took the set of verbs that had one of these containee nouns as a direct object and compared this set to the set of verbs that had the container noun *tape* as a direct object in the corpus. According to our hypothesis, verbs applying to the telic role should appear in the intersection of these two sets (as a result of the alternation), while those applying to the formal role will appear in the set difference {verbs with containers as direct object}—{verbs with containees as direct object}. The difference operation should serve to remove any verbs that co-occur with containee objects. Figure 3 shows the results of intersection and set difference for three container nouns *tape*, *disk*, and *directory*.

The results indicate that the container LCP is able to differentiate nouns with respect to their telic and formal qualia, for the nouns *tape* and *disk* but not for *directory*. The poor discrimination in the latter case can be attributed to the fact that a directory is a recursive container. A directory contains files, and a directory is itself a file. Therefore, verbs that apply to the formal role of directory are likely to apply to the formal role of objects contained in directories (such as other directories). This can be seen as a shortcoming of the container LCP for the task at hand, but may be a useful way of diagnosing when containers contain objects functionally similar to themselves.

The result of this corpus acquisition procedure is a kind of minimal faceted analysis for the noun *tape*, as illustrated below, showing only the qualia that are relevant to the discussion.[13]

$$\begin{bmatrix} \textbf{tape(x,y)} \\ \text{CONST} = \textbf{information(y);file(y)} \\ \text{FORMAL} = \textbf{mount(z,x);dismount(z,x)} \\ \text{TELIC} = \textbf{read(w,y);write(w,y);copy(w,y);contain(w,y)} \end{bmatrix}$$

13 Because the technique was sensitive to grammatical position of the object NP, the argument can be bound to the appropriate variable in the relation expressed in the qualia. It should be pointed out that these qualia values do not carry event place variables, since such discrimination was beyond the scope of this experiment.

What is interesting about the qualia values is how close they are to the concepts in the projective conclusion space of *tape*, as mentioned in Section 1.

To illustrate this procedure on another semantic category, consider the term *mouse* in its computer artifact sense. In our corpus, it appears in the object position of the verb *use* in a "use NP to" construction, as well as the object of the preposition *with* following a transitive verb and its object:

1. *use the mouse to set breakpoints*
2. *use the mouse anywhere*
3. *move a window with the mouse*
4. *click on it with the mouse ...*

These constructions are symptomatic of its role as an instrument; and the VP complement of *to* as well as the VP dominating the *with*-PP identify the telic predicates for the noun. Other verbs, for which *mouse* appears as a direct object are currently defaulted into the formal role, resulting in an entry for *mouse* as follows:

$$\begin{bmatrix} \textbf{mouse(x)} \\ \text{CONST} = \textbf{button(y)} \\ \text{FORMAL} = \textbf{physobj(x)} \\ \text{TELIC} = \textbf{set(x,breakpoint);move(x,window);click-on(x,z)} \end{bmatrix}$$

The above experiments have met with limited success, enough to warrant continuing our application of lexical semantic theory to knowledge acquisition from corpora, but not enough to remove the human from the loop. As they currently exist, the algorithms described here can be used as tools to help the knowledge engineer extract useful information from on-line textual sources, and in some applications (e.g., a "related terms" thesaurus for full text information retrieval) may provide a useful way to heuristically organize sublanguage terminology when human resources are unavailable.

4. Semantic Type Induction from Syntactic Forms

The purpose of the research described in this section is to experiment with the automatic acquisition of semantic tags for words in a sublanguage, tags well beyond that available from the seeding of MRDs. The identification of semantic tags is the result of type coercion on known syntactic forms, to induce a semantic feature, such as [+event] or [+object].

4.1 Coercive Environments in Corpora

A pervasive example of type coercion is seen in the complements of aspectual verbs such as *begin* and *finish*, and verbs such as *enjoy*. That is, in sentences such as "John began the book," the normal complement expected is an action or event of some sort, most often expressed by a gerundive or infinitival phrase: "John began reading the book," "John began to read the book." In Pustejovsky (1991) it was argued that in such cases, the verb need not have multiple subcategorizations, but only one *deep semantic type*, in this case, an event. Thus, the verb coerces its complement (e.g. "the book") into an event related to that object. Such information can be represented by means of a representational schema called *qualia structure*, which, among other things, specifies the relations associated with objects.

count	verb	object
205	*begin*	*career*
176	*begin*	*day*
159	*begin*	*work*
140	*begin*	*talk*
120	*begin*	*campaign*
113	*begin*	*investigation*
106	*begin*	*process*
92	*begin*	*program*
85	*begin*	*operation*
85	*begin*	*negotiation*
66	*begin*	*strike*
64	*begin*	*production*
59	*begin*	*meeting*
59	*begin*	*term*
50	*begin*	*visit*
45	*begin*	*test*
39	*begin*	*construction*
31	*begin*	*debate*
29	*begin*	*trial*

Figure 4
Counts for objects of *begin/V*.

In related work being carried out with Mats Rooth of the University of Stuttgart, we are exploring what the range of coercion types is, and what environments they may appear in, as discovered in corpora. Some of our initial data suggest that the hypothesis of deep semantic selection may in fact be correct, as well as indicating what the nature of the coercion rules may be. Using techniques described in Church and Hindle (1990), Church and Hanks (1990), and Hindle and Rooth (1991), Figure 4 shows some examples of the most frequent V-O pairs from the AP corpus.

Corpus studies confirm similar results for "weakly intensional contexts" such as the complement of coercive verbs such as *veto*. These are interesting because regardless of the noun type appearing as complement, it is embedded within a semantic interpretation of "the proposal to," thereby clothing the complement within an intensional context. The examples in Figure 5 with the verb *veto* indicate two things: first, that such coercions are regular and pervasive in corpora; second, that almost anything can be vetoed, but that the most frequently occurring objects are closest to the type selected by the verb.

What these data show is that the highest count complement types match the type required by the verb; namely, that one vetoes a bill or proposal to do something, not the thing itself. These nouns can therefore be used with some predictive certainty for inducing the semantic type in coercive environments such as "veto the expedition." This work is still preliminary, however, and requires further examination (Pustejovsky and Rooth [unpublished]).

4.2 Induction of Semantic Relations from Syntactic Forms

In this section, we present another experiment indicating the feasibility of inducing semantic tags for lexical items from corpora.[14] Imagine being able to take the V-O pairs

14 This section presents an abridged version of material reported on in Pustejovsky (1992).

count	verb	object
303	*veto*	*bill*
84	*veto*	*legislation*
58	*veto*	*measure*
35	*veto*	*resolution*
21	*veto*	*law*
14	*veto*	*item*
12	*veto*	*decision*
9	*veto*	*proposal*
9	*veto*	*plan*
7	*veto*	*package*
6	*veto*	*increase*
5	*veto*	*sanction*
5	*veto*	*penalty*
4	*veto*	*notice*
4	*veto*	*idea*
4	*veto*	*appropriation*
4	*veto*	*mission*
4	*veto*	*attempt*
3	*veto*	*search*
3	*veto*	*cut*
3	*veto*	*deal*
1	*veto*	*expedition*

Figure 5
Counts for objects of *veto/V*.

such as those given in Section 4.1, and then applying semantic tags to the verbs that are appropriate to the role they play for that object (i.e., induction of the qualia roles for that noun). This is similar to the experiment reported on in Section 3. Here we apply a similar technique to a much larger corpus, in order to induce the *agentive* role for nouns; that is, the semantic predicate associated with bringing about the object.

In this example we look at the behavior of noun phrases and the prepositional phrases that follow them. In particular, we look at the co-occurrence of nominals with *between, with,* and *to*. Table 1 shows results of the conflating noun plus preposition patterns. The percentage shown indicates the ratio of the particular collocation to the key word. Mutual information (MI) statistics for the two words in collocation are also shown. What these results indicate is that induction of semantic type from conflating syntactic patterns is possible. Based on the semantic types for these prepositions, the syntactic evidence suggests that there is an equivalence class where each preposition makes reference to a symmetric relation between the arguments in the following two patterns:

- `Z with y` = $\lambda R_Z \lambda x \exists y [R_Z(x,y) \wedge R_Z(y,x)]$
- `Z between x and y` = $\lambda R_Z \exists x,y [R_Z(x,y) \wedge R_Z(y,x)]$

We then take these results and, for those nouns where the association ratios for `N with` and `N between` are similar, we pair them with the set of verbs governing these "NP PP" combinations in corpus, effectively partitioning the original V-O set into [+agentive] predicates and [−agentive] predicates.

These are semantic n-grams rather than direct interpretations of the prepositions.

Table 1
Mutual information for noun + preposition patterns.

Word	Word + *to* (%)/MI	Word + *with* (%)/MI	Word + *between* (%)/MI	Word	Word + *to* (%)/MI	Word + *with* (%)/MI	Word + *between* (%)/MI
agreement	.117 1.512	.159 3.423	.028 3.954	expansion	.013 −.666	.007 .381	0 *n/a*
announcement	.010 −.918	.003 −.409	0 *n/a*	impasse	0 *n/a*	.064 2.520	.096 5.192
barrier	.215 2.117	0 *n/a*	.030 4.046	interactions	0 *n/a*	0 *n/a*	.250 6.141
competition	.019 −.269	.028 1.701	.021 3.666	market	.013 −.637	.006 .240	.000 −.500
confrontation	.029 .141	.283 4.000	.074 4.932	range	.005 −1.533	.002 −.618	.020 3.663
contest	.052 .715	.052 2.323	.039 4.301	relations	.009 −1.017	.217 3.736	.103 5.254
contract	.066 .947	.060 2.463	.002 1.701	settlement	.013 −.626	.091 2.868	.012 3.142
deal	.028 .086	.193 3.616	.004 2.015	talks	.029 .138	.218 3.740	.030 4.043
dialogue	0 *n/a*	.326 4.140	.152 5.644	venture	.032 .226	.105 3.008	.035 4.185
difference	.017 −.410	.009 .638	.348 6.474	war	.010 −.937	.041 2.079	.015 3.372

What these expressions in effect indicate is the range of semantic environments they will appear in. That is, in sentences like those in Example 16, the force of the relational nouns *agreement* and *talks* is that they are unsaturated for the predicate bringing about this relation. In 17, on the other hand, the NPs headed by *agreement* and *talks* are saturated in this respect.

Example 16
a. John made an agreement with Mary.
b. Apple opened talks with IBM.

Example 17
a. This is an agreement between John and Mary.
b. Those were the first talks between Apple and IBM.

If our hypothesis is correct, we expect that verbs governing nominals collocated with a *with*-phrase will be mostly those predicates referring to the agentive quale of the nominal. This is because the *with*-phrase is unsaturated as a predicate, and acts to

count	verb	object
19	*form*	*venture*
3	*announce*	*venture*
3	*enter*	*venture*
2	*discuss*	*venture*
1	*be*	*venture*
1	*abandon*	*venture*
1	*begin*	*venture*
1	*complete*	*venture*
1	*negotiate*	*venture*
1	*start*	*venture*
1	*expect*	*venture*

Figure 6
Verb-object pairs with prep = with.

identify the agent of the verb as its argument (cf. Nilsen (1973)). This is confirmed by our data, shown in Figure 6.

Conversely, verbs governing nominals collocating with a *between*-phrase will not refer to the agentive since the phrase is saturated already. Indeed, the only verb occurring in this position with any frequency is the copula *be*, namely with the following counts: `12 be/V venture/O`. Thus, weak semantic types can be induced on the basis of syntactic behavior.

There is a growing literature on corpus-based acquisition and tuning (Smadja 1991a; Zernik and Jacobs 1991; Brent 1991; as well as Grishman and Sterling 1992). We share with these researchers a general dependence on well-behaved collocational patterns and distributional structures. Probably the main distinguishing feature of our approach is its reliance on a fairly well studied semantic framework to aid and guide the semantic induction process itself, whether it involves selectional restrictions or semantic types.

5. Lexical Presuppositions and Preferences

In the previous section we presented algorithms for extracting collocational information from corpora, in order to supplement and fine-tune the lexical structures seeded by a machine-readable dictionary. In this section we demonstrate that, in addition to conventional lexical semantic relations, it is also possible to acquire information concerning lexical presuppositions and preferences from corpora, when analyzed with the appropriate semantic tools. In particular, we will discuss a phenomenon we call *discourse polarity*, and how corpus-based experiments provide clues toward the representation of this phenomenon, as well as information on preference relations.

As we have seen, providing a representational system for lexical semantic relations is a nontrivial task. Representing presuppositional information, however, is even more daunting. Nevertheless, there are some systematic semantic generalizations associated with such subtle lexical inferences. To illustrate this, consider the following examples taken from the *Wall Street Journal* Corpus, involving the verb *insist*.

Example 18
But Mr. Fourtou *insisted* that the restructuring plans hadn't played a role in his decision.

Example 19
But so far, the majority is *insisting* that a daily paper in the home is an essential educational resource that Mr. Oshry must have, like it or not.

Example 20
But Mr. Nishi *insists* there is a common theme to his scattered projects: to improve and spread personal computers.

Example 21
"Mister, Djemaa is a crazy place for you," *insists* the first of many young men, clutching a visitor's sleeve.

Example 22
But the BNL sources yesterday *insisted* that the head office was aware of only a small portion of the credits to Iraq made by Atlanta.

Example 23
Mr. Smale, who ordinarily *insists* on a test market before a national roll-out, told the team to go ahead—although he said he was skeptical that Pringle's could survive, Mr. Tucker says.

Example 24
The Cantonese *insist* that their fish be "fresh," though one whiff of Hong Kong harbor and the visitor may yearn for something shipped from distant seas.

Example 25
Money isn't the issue, Mr. Bush *insists*.

From analyzing these and similar data, a pattern emerges concerning the use of verbs like *insist* in discourse; namely, the co-occurrence with discourse markers denoting negative affect, such as *although* and *but*, as well as literal negatives, e.g., *no* and *not*. This is reminiscent of the behavior of *negative polarity items* such as *any more* and *at all*. Such lexical items occur only in the context of negatives within a certain structural configuration.[15] In a similar way, verbs such as *insist* seem to require an overt or implicit negation within the immediate discourse context, rather than within the clause. For this reason, we will call such verbs *discourse polarity items*.

For our purposes, the significance of such data is twofold: first, experiments on corpora can test and confirm linguistic intuitions concerning a subtle semantic judgment; second, if such knowledge is in fact so systematic, then it must be at least partially represented in the lexical semantics of the verb.

To test whether the intuitions supported by the above data could be confirmed in corpora, Bergler (1991) derived the statistical co-occurrence of *insist* with discourse polarity markers in the 7 million-word corpus of *Wall Street Journal* articles. She derived the statistics reported in Figure 7.

Let us assume, on the basis of this preliminary data[16] presented in Bergler (1992) that these verbs in fact do behave as discourse polarity items. The question then

15 There is a rich literature on this topic. For discussion see Ladusaw (1980) and Linebarger (1980).
16 Overlap between the categories occurs in less than 35 cases.

Keywords	Count	Comments
insist	586	occurrences throughout the corpus
insist on	109	these have been cleaned by hand and are actually occurrences of the idiom *insist on* rather than accidental co-occurrences.
insist & but	117	occurrences of both *insist* and *but* in the same sentence
insist & negation	186	includes *not* and *n't*
insist & subjunctive	159	includes *would, could, should,* and *be*

Figure 7
Negative markers with *insist* in WSJC .

immediately arises as to how we represent this type of knowledge. Using the language of the qualia structure discussed above, we can make explicit reference to the polarity behavior, in the following informal but intuitive representation for the verb *insist*.[17]

$$\left[\begin{array}{l} \textbf{insist(x:ind,y:prop)} \\ \text{FORMAL} = \text{REPORTING-VERB-LCP} \\ \text{TELIC} = \textbf{say(x,true(y)) \& presupposed}(\psi) \textbf{ \& y} = \neg\psi \end{array}\right]$$

This entry states that in the REPORTING–VERB sense of the word, *insist* is a relation between an individual and a statement that is the negation of a proposition, ψ, presupposed in the context of the utterance. As argued in Pustejovsky (1991) and Miller and Fellbaum (1991), such simple oppositional predicates form a central part of our lexicalization of concepts. Semantically motivated collocations such as these extracted from large corpora can provide presuppositional information for words that would otherwise be missing from the lexical semantics of an entry. While full automatic extraction of semantic collocations is not yet feasible, some recent research in related areas is promising.

Hindle (1990) reports interesting results of this kind based on literal collocations, where he parses the corpus (Hindle 1983) into predicate–argument structures and applies a *mutual information* measure (Fano 1961; Magerman and Marcus 1990) to weigh the association between the predicate and each of its arguments. For example, as a list of the most frequent objects for the verb *drink* in his corpus, Hindle found *beer, tea, Pepsi,* and *champagne.* Based on the distributional hypothesis that the degree of shared contexts is a similarity measure for words, he develops a similarity metric for nouns based on their substitutability in certain verb contexts. Hindle thus finds sets of semantically similar nouns based on syntactic co-occurrence data. The sets he extracts are promising; for example, the ten most similar nouns to *treaty* in his corpus are *agreement, plan, constitution, contract, proposal, accord, amendment, rule, law,* and *legislation.*

This work is very close in spirit to our own investigation here; the emphasis on syntactic co-occurrence enables Hindle to extract his similarity lists automatically; they

17 For illustration, we use an abbreviated version of the lexical entries under discussion, highlighting only certain qualia for the verbs. For the most recent representation of verbal semantics in this framework, see Pustejovsky (1993).

are therefore easy to compile for different corpora, different sublanguages, etc. Here we are attempting to use these techniques together with a model of lexical meaning, to capture deeper lexical semantic collocations; e.g., the generalization that the list of objects occurring for the word *drink* contains only *liquids*.

In the final part of this section, we turn to how the analysis of corpora can provide lexical semantic preferences for verb selection. As discussed above, there is a growing body of research on deriving collocations from corpora (cf. Church and Hanks 1990; Klavans, Chodorow, and Wacholder 1990; Wilks et al. 1993; Smadja 1991a, 1991b; Calzolari and Bindi 1990). Here we employ the tools of semantic analysis from Section 1 to examine the behavior of *metonymy* with reporting verbs. We will show, on the basis of corpus analysis, how verbs display marked differences in the ability to license metonymic operations over their arguments. Such information, we argue, is part of the preference semantics for a sublanguage, as automatically derived from corpus.

Metonymy can be seen as a case of "licensed violation" of selectional restrictions. For example, while the verb *announce* selects for a human subject, sentences like *The Phantasie Corporation announced third quarter losses* are not only an acceptable paraphrase of the selectionally correct form *Mr. Phantasie Jr. announced third quarter losses for Phantasie Corp*, but they are the *preferred* form in the *Wall Street Journal*). This is an example of subject *type coercion*, as discussed in Section 1. For example, the qualia structure for a noun such as *corporation* might be represented as below:

$$\begin{bmatrix} \textbf{corporation(x)} \\ \text{CONST} = \textbf{group(y),spokesperson(w),executive(z)} \\ \text{FORMAL} = \textbf{organization(x)} \\ \text{TELIC} = \textbf{execute(z,decisions)} \\ \text{AGENTIVE} = \textbf{incorporate(y,x)} \end{bmatrix}$$

The metonymic extension in this example is straightforward: a spokesman, executive, or otherwise legitimate representative "speaking for" a company or institution can be metonymically replaced by that company or institution.[18]

We find that this type of metonymic extension for the subject is natural and indeed very frequent with *reporting verbs* Bergler (1991), such as *announce, report, release,* and *claim,* while it is in general not possible with other verbs selecting human subjects, e.g., the verbs of contemplation (such as *contemplate, consider,* and *think*). However, there are subtle differences in the occurrence of such metonymies for the different members of the same semantic verb class that arise from corpus analysis.

A *reporting verb* is an utterance verb that is used to relate the words of a source. In a careful study of seven reporting verbs on a 250,000-word corpus of *Time* magazine articles from 1963, we found that the preference for different metonymic extensions varies considerably within this field (Bergler 1991). Figure 8 shows the findings for the words *insist, deny, admit, claim, announce, said,* and *told* for two metonymic extensions, namely where a *group* stands for an individual (*Analysts said* ...) and where a *company* or other *institution* stands for the individual (IBM announced ...).[19]

The difference in patterns of metonymic behavior is quite striking: semantically similar verbs seem to pattern similarly over all three categories; *admit, insist,* and *deny* show a closer resemblance to each other than to any of the others, while *said* and

18 Note, however, that the metonymic extension is not quite as simple as extending from any *employee* to the whole company or institution, but that a form of *legitimation* has to be involved.) For more detail see Bergler (1992).

19 The data for Figure 8 have been screened to ensure that only occurrences that constitute reporting contexts were used.

	person	group	instit.	other
admit	64%	19%	14%	2%
deny	59%	11%	19%	11%
insist	57%	24%	16%	3%
announce	51%	10%	31%	8%
claim	35%	21%	38%	6%
said	83%	6%	4%	8%
told	69%	7%	8%	16%

Figure 8
Preference for different metonymies in subject position.

	person	group	institution	other
WSJ	49%	15%	34%	2%
TIME	83%	6%	4%	8%

Figure 9
Preference for metonymies for *said* in a 160,000-word fragment of the *Wall Street Journal* corpus.

told form a category by themselves. There may be a purely semantic explanation why *said* and *told* seem not to prefer the metonymic use in subject position; e.g., perhaps these verbs relate more closely to the act of uttering, or perhaps they are too informal, stylistically. Evidence from other corpora, however, suggests that such information is accurately characterized as lexical preference. An initial experiment on a subset of the *Wall Street Journal* Corpus, for example, shows that *said* has a quite different metonymic distribution there, reported in Figure 9.

In this corpus we discovered that subject selection for an individual person appeared in only 50% of the sentences, while a company/institution appeared in 34% of the cases. This difference could either be attributed to a difference in *style* between *Time* magazine and the *Wall Street Journal* or perhaps to a difference in general usage between 1963 and 1989. The statistics presented here can of course not determine the reason for the difference, but rather help establish the lexical semantic preferences that exist in a certain corpus and sublanguage.

An important question related to the extraction of preference information is what the corpus should be. Recent effort has been spent constructing *balanced* corpora, containing text from different styles and sources, such as novels, newspaper texts, scientific journal articles, etc. The assumption is of course that given a *representative* mix of samples of language use, we can extract the general properties and usage of words. But if we gain access to sophisticated automatic corpus analysis tools such as those

discussed above, and indeed if we have specialized algorithms for sublanguage extraction, then homogeneous corpora might provide better data. The few examples of lexical preference mentioned in this section might not tell us anything conclusive for the definitive usage of a word such as *said*, if there even exists such a notion. Nevertheless the statistics provide an important tool for text analysis within the corpus from which they are derived. Because we can systematically capture the violation of selectional restrictions (as semantically predicted), there is no need for a text analysis system to perform extensive commonsense inferencing. Thus, such presupposition and preference statistics are vital to efficient processing of real text.

6. Summary and Discussion

In this paper we have presented a particularly directed program of research for how text corpora can contribute to linguistics and computational linguistics. We first presented a representation language for lexical knowledge, the generative lexicon, and demonstrated how it facilitates the structuring of lexical relations among words, looking in particular at the problems of metonymy and polysemy.

Such a framework for lexical knowledge suggests that there are richer relationships among words in text beyond that of simple co-occurrence that can be extracted automatically. The work suggests how linguistic phenomena such as metonymy and polysemy might be exploited for knowledge acquisition for lexical items. Unlike purely statistical collocational analyses, the framework of a semantic theory allows the automatic construction of predictions about deeper semantic relationships among words appearing in collocational systems.

We illustrated the approach for the acquisition of lexical information for several classes of nominals, and how such techniques can fine-tune the lexical structures acquired from an initial seeding of a machine-readable dictionary. In addition to conventional lexical semantic relations, we then showed how information concerning lexical presuppositions and preference relations can also be acquired from corpora, when analyzed with the appropriate semantic tools.

In conclusion, we feel that the application of computational resources to the analysis of text corpora has and will continue to have a profound effect on the direction of linguistic and computational linguistic research. Unlike previous attempts at corpus research, the current focus is supported and guided by theoretical tools, and not merely statistical techniques. We should furthermore welcome the ability to expand the data set used for the confirmation of linguistic hypotheses. At the same time, we must remember that statistical results themselves reveal nothing, and require careful and systematic interpretation by the investigator to become linguistic data.

Acknowledgments

This research was supported by DARPA contract MDA904-91-C-9328. We would like to thank Scott Waterman for his assistance in preparing the statistics. We would also like to thank Mats Rooth, Scott Waterman, and four anonymous reviewers for useful comments and discussion.

References

Ahlswede, T., and Evens, M. (1988). "Generating a relational lexicon from a machine-readable dictionary." *International Journal of Lexicography*, 1(3), 214–237.

Alshawi, H. (1987). "Processing dictionary definitions with phrasal pattern hierarchies." *Computational Linguistics*, 13, 3–4.

Alshawi, H.; Boguraev, B.; and Briscoe, T. (1985). "Towards a dictionary support environment for real time parsing." In *Proceedings, European Conference on Computational Linguistics*. Geneva, Switzerland.

Amsler, R. A. (1980). *The structure of the Merriam-Webster Pocket Dictionary*. Doctoral dissertation, University of Texas.

Amsler, R. A. (1989). "Third generation computational lexicology." In *Proceedings, First International Lexical Acquisition Workshop*. Detroit, Michigan, August 1989.

Amsler, R. A., and White, J. S. (1979). "Development of a computational methodology for deriving natural language semantic structures via analysis of machine-readable dictionaries." NSF Technical Report MCS77-01315.

Anick, P. (1992). "Lexicon assisted information retrieval for the help-desk." In *Proceedings, IEEE CAIA-92 Workshop on AI and Help-Desks*. Monterey, California.

Anick, P.; Brennan, J.; Flynn, R.; Hanssen, D.; Alvey, B.; and Robbins, J. (1989). "A direct manipulation interface for Boolean information retrieval via natural language query." In *Proceedings, SIGIR '89*.

Anick, P., and Pustejovsky, J. (1990). "An application of lexical semantics to knowledge acquisition from corpora." In *Proceedings, 13th International Conference of Computational Linguistics*. Helsinki, Finland.

Atkins, B. T. (1991). "Building a lexicon: Reconciling anisomorphic sense differentiations in machine-readable dictionaries." *International Journal of Lexicography*.

Atkins, B. T., and Levin, B. (1991). "Admitting impediments." In *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*, edited by U. Zernik. LEA.

Bergler, S. (1991). "The semantics of collocational patterns for reporting verbs." In *Proceedings, Fifth Conference of the European Chapter of the Association for Computational Linguistics*. Berlin, Germany, April 1991.

Bergler, S. (1992). "Evidential analysis of reported speech." Doctoral dissertation, Brandeis University.

Bierwisch, M. (1983). "Semantische und konzeptuelle Repräsentationen lexikalischer Einheiten." In *Untersuchungen zur Semantik*, edited by R. Ruzicka and W. Motsch. Akademische-Verlag.

Binot, J.-L., and Jensen, K. (1987). "A semantic expert using an online standard dictionary." In *Proceedings, 10th International Joint Conference on Artificial Intelligence (IJCAI-87)*. Milan, Italy, 709–714.

Boguraev, B. (1979). "Automatic resolution of linguistic ambiguities." Technical Report No. 11, University of Cambridge Computer Laboratory, Cambridge, U.K.

Boguraev, B. (1991). "Building a lexicon: The contribution of computers." *International Journal of Lexicography*, 4(3).

Boguraev, B., and Briscoe, T. (1989). "Introduction." In *Computational Lexicography for Natural Language Processing*, edited by B. Boguraev and T. Briscoe. Longman Group UK.

Boguraev, B., and Briscoe, T. (1987). "Large lexicons for natural language processing: Exploring the grammar coding system of LDOCE." *Computational Linguistics*, 13.

Boguraev, B.; Byrd, R.; Klavans, J.; and Neff, M. (1989). "From machine readable dictionaries to a lexical knowledge base." In *Proceedings, First International Lexical Acquisition Workshop*. Detroit, Michigan, August 1989.

Boguraev, B., and Pustejovsky, J. (1990). "Lexical ambiguity and the role of knowledge representation in lexicon design." In *Proceedings, 13th International Conference of Computational Linguistics*. Helsinki, Finland, August 1990.

Brent, M. (1991). "Automatics semantic classification of verbs from their syntactic contexts: An implemented classifier for stativity." In *Proceedings, Fifth Conference of the European Chapter of the Association for Computational Linguistics*. Berlin, Germany, April 1991.

Briscoe, E.; Copestake, A.; and Boguraev, B. (1990). "Enjoy the paper: Lexical semantics via lexicology." In *Proceedings, 13th International Conference on Computational Linguistics*. Helsinki, Finland, 42–47.

Byrd, R.; Calzolari, N.; Chodorow, M.; Klavans, J.; Neff, M.; and Rizk, O. (1987). "Tools and methods for computational lexicology." *Computational Linguistics*, 13(3–4), 219–240.

Byrd, R. (1989). "Discovering relationships among word senses." In *Proceedings, Fifth Annual Conference of the UW Centre for the New Oxford English Dictionary*. Oxford, U.K., 67–80.

Calzolari, N. (1984). "Detecting patterns in a lexical database." In *Proceedings, Seventh International Conference on Computational Linguistics (COLING-84)*. Stanford, California.

Calzolari, N., and Bindi, R. (1990). "Acquisition of lexical information from a large textual Italian corpus." In *Proceedings, 13th International Conference on Computational Linguistics (COLING-90)*. Helsinki, Finland.

Cardelli, L., and Wegner, P. (1985). "On understanding types, data abstraction, and polymorphism." *ACM Computing Surveys*, 17(4), 471–522.

Chodorow, M.; Byrd, R.; and Heidorn, G.

(1985). "Extracting semantic hierarchies from a large on-line dictionary." In *Proceedings, 23rd Annual Meeting of the ACL*. Chicago, Illinois, 299–304.

Church, K. (1988). "A stochastic parts program and noun phrase parser for unrestricted text." In *Proceedings, Second Conference on Applied Natural Language Processing*. Austin, Texas.

Church, K., and Hanks, P. (1990). "Word association norms, mutual information, and lexicography." *Computational Linguistics*, 16(1), 22–29.

Church, K., and Hindle, D. (1990). "Collocational constraints and corpus-based linguistics." In *Working Notes of the AAAI Symposium: Text-Based Intelligent Systems*. Stanford, California.

Copestake, Ann (in press). "Defaults in the LKB." In *Default Inheritance in the Lexicon*, edited by T. Briscoe and A. Copestake. Cambridge University Press.

Copestake, A., and Briscoe, E. (1992). "Lexical operations in a unification-based framework." In *Lexical Semantics and Knowledge Representation*, edited by J. Pustejovsky and S. Bergler. Springer Verlag.

Cowie, J.; Guthrie, L.; and Pustejovsky, J. (1992). "Description of the MUCBRUCE system as used for MUC-4." In *Proceedings, Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufmann.

Croft, W. B. (1989). "Automatic indexing." In *Indexing: The State of Our Knowledge and the State of Our Ignorance*, edited by B. H. Weinberg, 87–100. Learned Information, Inc.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Cruse, D. A. (1992). "Polysemy and related phenomena from a cognitive linguistic viewpoint." In *Proceedings, Second Toulouse Workshop on Lexical Semantics*, edited by P. Saint-Dizier and E. Viegas. Toulouse, France.

Debili, F.; Fluhr, C.; and Radasoa, P. (1988). "About reformulation in full-text IRS." In *Proceedings, RIAO-88*, 343–357.

Evans, R., and Gazdar, G., editors. (1990). "The DATR papers: February 1990." Cognitive Science Research Paper CSRP 139, School of Cognitive and Computing Sciences, University of Sussex.

Evens, M. (1987). *Relational Models of the Lexicon*. Cambridge University Press.

Fano, R. (1961). *Transmission of Information: A Statistical Theory of Communications*. MIT Press.

Fillmore, C. (1968). "The case for case." In *Universals in Linguistics Theory*, edited by E. Bach and R. Harms. Holt, Rinehart and Winston.

Grishman, R.; Hirschman, L.; and Nhan, N. T. (1986). "Discovery procedures for sublanguage selectional patterns: Initial experiments." *Computational Linguistics*, 12(3), 205–215.

Grishman, R., and Sterling, J. (1992). "Acquisition of selectional patterns." In *Proceedings, 14th International Conference on Computational Linguistics (COLING-92)*. Nantes, France, July 1992.

Guthrie, L.; Slator, B.; Wilks, Y.; and Bruce, R. (1990). "Is there content in Empty Heads?" In *Proceedings, 13th International Conference of Computational Linguistics (COLING-90)*. Helsinki, Finland.

Hindle, D. 1983. "Deterministic parsing of syntactic non-fluencies." In *Proceedings, 21st Annual Meeting of the Association for Computational Linguistics*, Cambridge, Massachusetts, June 1983, 123–128.

Hindle, D. (1990). "Noun classification from predicate-argument structures." In *Proceedings, 28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh, Pennsylvania, June 1990, 268–275.

Hindle, D., and Rooth, M. (1991). "Structural ambiguity and lexical relations." In *Proceedings of the ACL*.

Jacobs, P. (1991). "Making sense of lexical acquisition." In *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*, edited by U. Zernik, LEA.

Klavans, J.; Chodorow, M.; and Wacholder, N. (1990). "From dictionary to knowledge base via taxonomy." In *Proceedings, Sixth Conference of the UW Centre for the New OED*. Waterloo, 110–132.

Krovetz, R., and Croft, W. B. (1989). "Word sense disambiguation using machine-readable dictionaries." In *Proceedings, SIGIR*. 127–136.

Ladusaw, W. (1980). *Polarity Sensitivity as Inherent Scope Relations*. Indiana University Linguistics Club.

Lakoff, G. (1968). "Instrumental adverbs and the concept of deep structure." *Foundations of Language*, 4, 4–29.

Lakoff, G. (1970). *Irregularity in Syntax*. Holt, Rinehart, and Winston.

Lakoff, G. (1987). *Women, Fire, and Dangerous Objects*. University of Chicago Press.

Linebarger, M. (1980). "The grammar of negative polarity." Doctoral dissertation, MIT, Cambridge MA.

Maarek, Y. S., and Smadja, F. Z. (1989). "Full text indexing based on lexical relations, an application: Software libraries." In *Proceedings, SIGIR*. 127–136.

Markowitz, J.; Ahlswede, T.; and Evens, M. (1986). "Semantically significant patterns in dictionary definitions." In *Proceedings, 24th Annual Meeting of the Association for Computational Linguistics.* New York, New York, 112–119.

Magerman, D., and Marcus, M. (1990). "Parsing a natural language using mutual information statistics." In *Proceedings, Eighth National Conference on Artificial Intelligence (AAAI-90).* Boston, Massachusetts.

Martin, J. (1990). *A Computational Model of Metaphor Interpretation.* Academic Press.

Mel'čuk, I. (1988). *Dependency Syntax.* SUNY Press.

Miller, G., and Fellbaum, C. (1991). "Verbs in WordNet." *Cognition.*

Nakamura, J., and Nagao, M. (1988). "Extraction of semantic information from an ordinary English dictionary and its evaluation." In *Proceedings, COLING-88.* Budapest, Hungary, 459–464.

Nilsen, D. L. F. (1973). *The Instrumental Case in English: Syntactic and Semantic Considerations.* Mouton.

Nirenburg, S., and Nirenburg, I. (1988). "A framework for lexical selection in natural language generation." In *Proceedings, COLING-88.* Budapest, Hungary.

Nunberg, G. (1978). *The Pragmatics of Reference.* Indiana University Linguistics Club.

Procter, P.; Ilson, R. F.; and Ayto, J. (1978). *Longman Dictionary of Contemporary English.* Longman Group Limited.

Pustejovsky, J. (1989). "Issues in computational lexical semantics." In *Proceedings, Fourth Conference of the European Chapter of the ACL.* Manchester, England, April 1989.

Pustejovsky, J. (1991). "The generative lexicon." *Computational Linguistics,* 17(4), 409–441.

Pustejovsky, J. (1992). "The acquisition of lexical semantic knowledge from large corpora." In *Proceedings, DARPA Spoken and Written Language Workshop.* Morgan Kaufmann.

Pustejovsky, J. (1993). "Linguistic constraints on type coercion." In *Computational Lexical Semantics,* edited by P. Saint-Dizier and E. Viegas. Cambridge University Press.

Pustejovsky, J. (in press). *The Generative Lexicon: A Theory of Computational Lexical Semantics.* MIT Press.

Pustejovsky, J., and Anick, P. (1988). "The semantic interpretation of nominals." In *Proceedings, 12th International Conference of Computational Linguistics.* Budapest, Hungary, August 1988.

Pustejovsky, J., and Bergler, S. (1987). "The acquisition of conceptual structure for the lexicon." In *Proceedings, Sixth National Conference on Artificial Intelligence.* Seattle, Washington.

Pustejovsky, J., and Boguraev, B. (1993). "Lexical knowledge representation and natural language processing." *Artificial Intelligence.*

Pustejovsky, J., and Rooth, M. (unpublished). "Type coercive environments in corpora."

Pustejovsky, J.; Waterman, S.; Cowie, J.; and Stein, G. (1992). "Overview of the *DIDEROT* system for the Tipster text extraction project." In *Proceedings, DARPA TIPSTER 12-Month Evaluation.* San Diego, California, September 1992.

Rau, L., and Jacobs, P. (1988). "Integrating top-down and bottom-up strategies in a text processing system." In *Proceedings, Second Conference on Applied Natural Language Processing,* Austin Texas, February 1988, 129–135.

Russell, G.; Ballim, A.; Carroll, J.; and Warwick-Armstrong, S. (1992). "A practical approach to multiple default inheritance for unification-based lexicons." *Computational Linguistics,* 18(3), 311–337.

Slator, B. M. (1988). "Constructing contextually organized lexical semantic knowledge-bases." In *Proceedings, Third Annual Rocky Mountain Conference on Artificial Intelligence.* Denver, Colorado, June 1988, 142–148.

Slator, B. M., and Wilks, Y. A. (1987). "Toward semantic structures from dictionary entries." In *Proceedings, Second Annual Rocky Mountain Conference on Artificial Intelligence.* Boulder, Colorado, 85–96.

Smadja, F. (1991a). "Macrocoding the lexicon with co-occurrence knowledge." In *Lexical Acquisition: Using On-Line Resources to Build a Lexicon,* edited by U. Zernik. Lawrence Erlbaum Associates.

Smadja, F. (1991b). "From n-grams to collocations: an evaluation of xtract." In *Proceedings, 29th Annual Meeting of the Association for Computational Linguistics.* Berkeley, California, June 1991, 279–284.

Sparck Jones, K., editor. (1981). *Information Retrieval Experiments.* Butterworth.

Sparck Jones, K. (1986). *Synonymy and Semantic Classification.* Edinburgh Information Technology Series (EDITS). Edinburgh University Press.

Talmy, L. (1985). "Lexicalization patterns." In *Language Typology and Syntactic Description 3: Grammatical Categories and the*

Lexicon, edited by T. Shopen, 57–149. Cambridge University Press.

Touretzky, D. S. (1986). *The Mathematics of Inheritance Systems*. Morgan Kaufmann.

Veronis, J., and Ide, N. (1991). "An assessment of semantic information automatically extracted from machine readable dictionaries." In *Proceedings, Fifth Conference of the European Chapter of the Association for Computational Linguistics*. Berlin, Germany, April 1991.

Vickery, B. C. (1975). *Classification and Indexing in Science*. Butterworth and Co.

Walker, D. E., and Amsler, R. A. (1986). "The use of machine-readable dictionaries in sublanguage analysis." In *Analyzing Language in Restricted Domains*, edited by R. Grishman and R. Kittredge. Lawrence Erlbaum Associates.

Vossen, P.; Meijs, W.; and den Broeder, M. (1989). "Meaning and structure in dictionary definitions." In *Computational Lexicography for Natural Language Processing*, edited by B. Boguraev and T. Briscoe. Longman.

Wang, Y.-C.; Vandendorpe, J.; and Evens, M. (1985). "Relational thesauri in information retrieval." *Journal of the American Society for Information Science*, 36, 15–27.

Wilensky, R.; Chin, D. N.; Luria, M.; Martin, J.; Mayfield, J.; and Wu, D. (1988). "The Berkeley UNIX consultant project." *Computational Linguistics*, 14(4), 35–84.

Wilks, Y. A. (1978). "Making preferences more active." *Artificial Intelligence*, 10, 75–97.

Wilks, Y. A.; Fass, D.; Guo, C.-M.; McDonald, J. E.; Plate, T.; and Slator, B. M. (1989). "A tractable machine dictionary as a resource for computational semantics." In *Computational Lexicography for Natural Language Processing*, edited by B. Boguraev and T. Briscoe, 193–228. Longman.

Wilks, Y.; Fass, D.; Guo, C.-M.; McDonald, J. E.; Plate, T.; and Slator, B. M. (1993). "Providing machine tractable dictionary tools." In *Semantics and the Lexicon*, edited by J. Pustejovsky. Kluwer Academic Publishers.

Zernik, U. (1989). "Lexicon acquisition: Learning from corpus by exploiting lexical categories." In *Proceedings, IJCAI-89*.